elsnews

9.2

The Newsletter of the European Network in Human Language Technologies Summer 2000

LREC 2000 – Resources for the Millennium?

***Dimitrios Kokkinakis**, Göteborg University*

Zappeion Megaron, central Athens

The second Conference on Language Resources and Evaluation (LREC) was hosted by the Institute of Language and Speech Processing (ILSP), at the prestigious exhibition hall of Zappeion Megaron in the centre of Athens, from 31 May to 2 June 2000. To accommodate the large number of contributors, events were structured into four thematic sessions. The four domains were: written area; spoken resources and evaluation; evaluation in the written area; and terminology. Each domain was covered by both regular papers (30 sessions) and some of the large number of posters (16 sessions). For the first time, I think, the posters successfully achieved first-class status as an integrated part of a conference. Eleven satellite pre- and post-conference workshops completed the picture of a successful venue, which illustrated the variety of issues that needed to be addressed and discussed, during 300

the following I will try to give a very brief, but necessarily subjective, survey of what I perceived as the major areas of emphasis and general trends in current work on Language Resources (LR) and Human Language Technology (HLT).

The conference proper began with a panel organised by Catherine Macleod, entitled ‘Resources for the Millennium’. This comprised a series of presentations in various flavours, emphasising key topics such as multilingual resources, public domain tools, and specialised corpora. In addition to the more ‘traditional’ themes, there were some which charted new directions within a familiar area: the standardisation of resources, this time in the XML era (eXtensible Markup Language). We have for quite some time been familiar with work on the Text Encoding Initiative (TEI), Corpus Encoding Standard (CES), and

Summer 2000

elsnet

Conference Report (cont.)

ELSNews is published at the School of Cognitive and Computing Sciences, University of Sussex. It is printed by the University of Sussex Print Unit.

Editor: Jenny Norris.

Editorial Team:
Geoffrey Sampson,
Steven Krauwer and
Brigitte Burger.

ISSN 1350-990X

© ELSNET 2000

FOR INFORMATION
Contributions to *ELSNews*, and corrections, should be sent to:
jennyn@cogs.sussex.acuk

Tel: +44 1273 678659
Fax: +44 1273 671320

Material for the next issue is due:
15 OCTOBER 2000

Mr. P. Efthimiou, Greek Minister of Education, addressing the conference at the opening session. Next to him, three members of the organising committee – from left to right: Prof. G. Carayannis, Director of ILSP; Prof. A. Zampolli, Director of ILC and President of ELRA; and Dr. K. Choukri, Chief Executive Officer of ELRA

time, however, there was a strong feeling that the standardisation of resources, and applications and tools that actually implement the emerging and powerful XML technology, are just around the corner: resources that will enable us to easily encode, manage, re-use, and explore complex textual, spoken, or multimodal data. Nancy Ide and her co-workers gave a good insight into the capabilities of the XML mechanisms that are most relevant to language engineering research, and we should expect a move towards it in the near future – both at the annotation level, with software support, and the product level.

As usual, a number of presentations reported progress in ongoing projects, such as the American National Corpus (ANC), or recently completed ones, such as the Semantic Information for Multifunctional Plurilingual Lexica (SIMPLE). The importance of aspects relating to the lexicon were emphasised during discussions of SIMPLE – and the lexicon was also at the centre of a considerable number (over 40) of contributions, many of which had SIMPLE as the core. Also within the same domain, Adam Kilgarriff informed the audience about some exciting research tasks in the direction signalled by the SENSEVAL (SENSe EVALuation) exercise, and the plans for a continuation with more languages and more rigorous evaluation criteria. Word sense disambiguation, cross-language encoding, and the acquisition and linking of semantic lexical resources and their applications, are some of the open issues that will be under continuing focus in future activities.

During the first LREC in Granada in 1998, issues were raised regarding the infrastructure dedicated to LR and Eastern European and other less common languages. The current research climate once more emphasises these issues. A workshop, poster and panel sessions, and several papers in the main conference were dedicated to the continuing discussion of the development of LR for such languages. These, and other similar activities, conveyed a strong impression that work has been rapidly advanced for many 'minority' languages, such as Basque, Czech, Croat, Maltese and Slovenian. Whether or not the funding agencies, both national and international, will continue to support the work achieved so far, remains to be seen.

The oral and software presentations provided evidence of the impact that HLT has had on the activities of leading companies in the field, such as Microsoft, IBM, L&H, and Motorola Labs – demonstrating, once again, how seriously they consider HLT as an integrated part of their activities. This is very positive. However, the question that struck me by the end of the conference was whether we are about to experience the centralisation of resources, making them almost unreachable for those of us on the outside, including those in academia. Time will show how we can find better formulas for co-operation, finance, and support for the future in this field.

The sessions were generally very interesting: however, in some of the panels, the audience felt that a stimulating discussion and debate between the panellists, and between the panellists and the audience, was lacking. A similar issue was raised by Nicoletta Calzolari at the summing-up session, where she also expressed her wish for more contributions to the discussion, particularly from the industrially-oriented delegates.

The social events included an excursion to the ancient site of Delphi and, inevitably, a presentation of the preparation and work done so far for the summer Olympics in 2004, with emphasis on the Hellenic heritage of the subject. The conference officially ended with an extraordinary gala dinner at the Hilton Hotel where Antonio Zampolli was once again acclaimed as the director of entertainment, and Christiane Fellbaum and Piek Vossen announced the international society for WordNet. George Carayannis and his staff have every reason to be proud of having organised such a fruitful and stimulating conference. By the way, was it the Microsoft team that won the ancient parody competition?

FOR INFORMATION

Dimitrios Kokkinakis is Research Fellow in Lexicography and NLP at Språkdata, in the Department of Swedish Language at Göteborg University in Sweden.

Email: svedk@svenska.gu.se

Web: http://svenska.gu.se/~svedk

Conference Report

The First International Conference on Natural Language Generation: INLG 2000 and Language Processing in Israel

***John Carroll**, University of Sussex*

John Carroll

After nine biennial international workshops on natural language generation, from 1982 to 1998, INLG has been reborn as a conference series. INLG 2000 took place June 12-16 in Mitzpe Ramon (Israel), an isolated location with stunning scenery in the middle of the Negev Desert. The conference programme included thirty main presentations, plus student presentations, invited talks, system demonstrations, and associated workshops (Analysis for Generation and Coherence in Generated Multimedia). Contributors' affiliations covered twelve countries, spanning North and South America, Europe, the Middle East, and Australasia.

A common strand of work reported at the conference was the re-use of resources for generation systems, with papers on generic system frameworks, wide-coverage lexicons, and shareable processing components. Another popular theme tackled the generation of utterances in the context of large-scale application systems, such as speech-to-speech translation, dynamically-created multimedia presentations, and multi-lingual and/or multi-document summarisation.

One of the highlights of the conference for me was a paper session and panel discussion on the evaluation of generation systems. Evaluation has not received as much attention from researchers in generation as in some other areas of speech and language processing, for example, information extraction and retrieval (with US DARPA-sponsored exercises), or indeed parsing (see *ELSNews* 7.3, July 1998). Some promising outline proposals were put forward; however, it is much harder to evaluate a particular generation module than, for instance, a parser, since the generation task is in general wider and less well-defined. Thus, in contrast to producing syntax trees for individual sentences, a generation system might take as input a semantic representation and produce a number of sentences, often with many alternative phrasings being equally acceptable. The evaluation issue will without doubt resurface at future generation conferences.

There is considerable activity in language processing research in Israel. Israel is not a member of the European Union, although by virtue of being an 'associated state', research there is eligible for EU funding in the Fourth and Fifth Framework Programmes (in common with Iceland and Norway, for example). Much Israeli academic research in this field is fairly accessible, appearing in mainstream international computational linguistics publications, and often concerns the processing of English.

Michael Elhadad, of Ben-Gurion University at Beer Sheva, was the INLG conference chair. Elhadad and his colleagues have developed FUF/SURGE, a back-end realisation module (for English) that is widely used in the generation community: other current work is focussing on applying generation techniques to automatic summarisation and intelligent document creation. Bar Ilan University, near Tel Aviv, is another major centre for language processing; Ido Dagan and co-workers have developed innovative statistical techniques and applied them to multilingual information access, text mining, the alignment of parallel texts in disparate languages, and shallow parsing.

At Technion – the Israel Institute of Technology, in Haifa – Nissim Francez and colleagues are working on natural language semantics and on formal issues in unification-based grammar processing. Alon Itai has been involved in research on morphological and word sense disambiguation of Hebrew. Hebrew, the main official language of Israel, is a highly inflected language, which may explain the fact that, to date, there has been comparatively little computational work on it above the morphological level. However, this may change since there are now extensive linguistic resources available: for instance Yaacov Choueka of Bar Ilan University has led a large team of linguists, computational linguists, and lexicographers in the development of a comprehensive and robust set of tools (lemmatisers, etc.) and data (dictionaries) for the processing of Hebrew.

FOR INFORMATION

John Carroll is an Advanced Research Fellow at the University of Sussex.

Email: johnca@cogssusx.ac.uk

Web: http://www.cogssusx.ac.uk/lab/nlp/carroll/

The conference programme and Postscript versions of the papers are online at http://www.csbgu.ac.il/~nlg2000/

Descriptions of the associated workshops are at
http://www.csbgu.ac.il/~nlg2000/ws1.html and
http://www.csbgu.ac.il/~nlg2000/ws3.html

Project Update

TRINDI Project Nears Conclusion

Robin Cooper, *Göteborg University*

Robin Cooper (photo by Anton Nijholt)

The EU Language Engineering research project TRINDI (Task Oriented Instructional Dialogue) concludes at the end of August 2000. The project, which is a collaboration between Göteborg University (scientific coordinator), Edinburgh University (financial coordinator), Universität des Saarlandes, SRI Cambridge, and Xerox Research Centre Europe, has tackled some fundamental problems in the design of dialogue systems. Its central focus has been on the representation of information about the state of the dialogue (e.g., what questions are under discussion, what obligations a dialogue participant has at a given stage of the dialogue) and how such information should be updated as the dialogue progresses. The project has taken a practical approach to these theoretical issues by implementing a toolbox, the TRINDIKIT, which allows users to experiment with different kinds of information states and rules for updating them.

The economic and societal impact of spoken language dialogue systems is vast and potentially far-reaching in the changes it may bring about in the information society. The emphasis is shifting from traditional computer interfaces, with screens and keyboards, to computational modules embedded in appliances, hand-held devices, and information systems. As size decreases and mobility increases, and computational elements become more and more ubiquitous in our daily life (from smart houses to internet access via mobile phone), spoken language dialogue is increasingly becoming the most attractive interface. There is potential here for a profound change in the way we interact with our machines. The current state of the art in dialogue technology does not yet provide for systems which have moderately complex behaviour and which can easily be ported to different domains. The dialogue system toolkits that exist (e.g., the Philips and Nuance systems) give quite restricted behaviour and nevertheless require a considerable effort to create a new dialogue system. Systems which have more complex behaviour, e.g., based on reasoning about plans, tend to require a large amount of work if they are to be ported to different domains or languages. This project aimed to find a middle ground by developing a theoretical approach based on information states, which is simple enough to allow rapid prototyping of experimental systems, and to create a framework in which it is easier to port from one domain to another and from one language to another.

The three main objectives listed in the project's work plan were to:

- analyse features of human-human task-oriented instructional dialogue which have to do with the way

participants change during the course of the dialogue

- examine how such features can be modified in order to simplify the task of facilitating human-machine interaction which is both natural and robust, even though it is more restricted than human-human dialogue

- build a computational model of information revision in task-oriented and instructional dialogue and instructional texts

We considered three domains in pursuing these objectives: autoroute (planning a route between two cities); travel planning (flight booking); and machine repair (Xerox machine manual). The dialogues we considered in this project were for the most part *information seeking* – where the system normally has a series of questions which the user must answer in order to enable the system to perform some task. Among other things, a successful flexible system needs to be able to recognise that it has received the answer to a question even if it has not yet asked the question. It must also be able to cope with incomplete answers, and with answers that give more information than was expected.

There were five main themes in the project:

Analysis of instructional dialogues and dialogue systems

Following the overall aim of the project to relate information states to dialogue moves, in a way that allows for practical implementation, we examined some instructional dialogues and texts (concentrating on autoroute and machine repair domains) and developed methods for coding for information states.

At the same time, we gave a brief survey of some existing dialogue systems, and attempted to develop a way of evaluating dialogue systems relevant to information states (the TRINDI Ticklist).

Computational model of dialogue dynamics

This theme represented the heart of our theoretical and implementational development. We developed a theoretical view of information state update in dialogue that would encompass alternative views – such as the Poesio-Traum view of conversational updates, based on micro-conversational events; DRT, Ginzburg's dialogue game board with Questions Under Discussion (QUD); and Conversational Game Theory as it has been developed at SRI Cambridge.

We also developed a toolkit environment (TRINDIKIT), in which it is possible to implement dialogue move engines and experiment with different kinds of information state and information update rules and algorithms. Our hope is that the TRINDIKIT will provide an environment for the rapid prototyping of dialogue systems based on general theoretical

Validation of the model in the scenarios

Validation was carried out by showing that the general theoretical and implementational tools we developed could be applied to particular domains and implementations (including text as a limit case of dialogue), and could account for various levels of interaction between user and machine. The implementations that were carried out as part of the validation are:

GoDiS – a simple implementation based on Ginzburg's dialogue game board, which shows how the notion of accommodation can be exploited

MIDAS – a DRT based implementation including aspects of the Poesio-Traum approach, which exploits first order theorem proving in dialogue updates

EDIS – a more direct implementation of the Poesio-Traum approach

IMDIS – a simple adaptation of GoDiS to deal with texts, which shows that a single system can be used both to generate text and to conduct a conversation on the content of the text.

SRI autoroute – a demonstration that the SRI autoroute system, based on conversational games, can be cast in terms of the TRINDIKIT

SRI robust – a demonstration of robust semantic processing

Information structure in dialogue dynamics

This theme was more exploratory than the preceding ones. It deals with ways of representing focus-ground articulation and parallelism in the kind of information states we have been using, and how prosody can be related to these information states. We have implemented a demonstration showing that there are simple strategies for predicting focus prosody (e.g., for parallel questions) from information states, without having to encode focus information as such overtly in the information state. This can be used to improve the performance of off-the-shelf text-to-speech systems connected to a dialogue system built with the TRINDIKIT.

Modelling robust processing of dialogue

A theme that we have pursued throughout all our work on the project is the relevance and nature of underspecified and robust processing to the kind of dialogue systems we are interested in.

Annie Zaenen (Xerox Research Centre Europe) and Staffan Larsson (Göteborg University) present joint work at the final TRINDI review

During the final phase of the project we intend to produce a single document that will present the results in a more accessible fashion for a general audience. At the same time we hope to get the TRINDIKIT into a form where it might be user friendly enough to be generally used by dialogue system developers interested in a 'plug and play' approach to dialogue systems based on information states.

FOR INFORMATION

Robin Cooper is Professor in Computational Linguistics at Göteborg University, and coordinator of TRINDI.

Email: trindi@ling.gu.se

For further information, together with deliverables and other references (including those in this report), demos, and a downloadable version of the TRINDIKIT:

Web: http://www.ling.gu.se/research/projects/trindi

Book Announcement and Offer

ELSNews Letters Page

ELSNews has received this topical letter from Adam Kilgarriff, at the University of Brighton's Information Technology Research Institute.

ELSNews invites responses to this letter, as well as readers' comments and opinions relating to any other material appearing in its pages. Your views – even informally expressed – on any matters related to human language technologies are always welcomed.

Send your contributions to the Editor, Jenny Norris, by 15 October for inclusion in the next issue.

Email: *jennyn@cogs.sussex.ac.uk*

Dear Jenny,

Anyone who has worked with corpora will be all too aware of differences between them. Depending on the differences between them, it may, or may not, be reasonable to expect results based on one corpus to be also valid for another. It may, or may not, be appropriate for a grammar, or parser, based on one to perform well on another. It may, or may not, be straightforward to port an application from a domain of the first text type to a domain of the second. Currently, characterisations of corpora are mostly textual. A corpus is described as, say, 'Wall Street Journal' or 'transcripts of business meetings' or 'foreign learners' essays (intermediate grade)'. It would be desirable to be able to place a new corpus in relation to existing ones, and to be able to quantify similarities and differences.

Allied to corpus-similarity is corpus-homogeneity. An understanding of homogeneity is a prerequisite to a measure of the similarity – it makes little sense to compare a corpus sampled across many genres, like the Brown, with a corpus of weather forecasts. You want to be able to say, simply, that the first is broad, the second, narrow.

Given the centrality of corpora to contemporary language engineering, it is remarkable how little research there has been to date on the question. Some of the most salient work is Doug Biber's, where the goal is explicitly to identify the differences between different text types. There has been continuing work in this paradigm, with mixed results.

Comparing corpora is a variant on comparing texts, and the field of text classification is a close relative, though the standard assumptions in text classification of a set of pre-defined categories, and that the interesting differences relate to subject area rather than genre, mean that much work is not salient. The TypTex system (Folch et al., 2000) demonstrates one text classification system that does not make these assumptions and which can be applied to comparing corpora.

The only well-understood measures are perplexity and cross-entropy, from information theory: these are used to assess homogeneity and similarity in language modelling, particularly for speech recognition (e.g, Roukos, 1996). However, it is an assumption that they are suitable measures, which match pre-theoretic judgements and which capture the aspects of homogeneity and similarity that are of interest for various language engineering purposes; and there is evidence that they do not match our intuitions well (Kilgarriff and Rose, 1998).

There are, of course, many ways in which two corpora will differ, and different kinds of differences will be relevant for different kinds of purposes. Thus, similarity such that a part-of-speech tagger developed for one corpus works well in the other may differ from similarity for machine translation. We currently lack a sophisticated vocabulary for talking about the different ways in which corpora differ, and how these might correlate with the costs of porting LE applications.

There will be a workshop on 'Comparing Corpora' in association with the ACL meeting in Hong Kong (the workshop will be on 7th or 8th October), where we anticipate taking these questions forward and working out improved ways of measuring corpus similarity and corpus homogeneity.

Yours sincerely,

Adam

Adam Kilgarriff's letter refers to the following:

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Biber, D. 1995. *Dimensions in Register Variation*. Cambridge University Press.

Folch, H. et al. 2000. TyPTex: Inductive typological text classification by multivariate statistical analysis for NLP systems tuning/evaluation. Proc. LREC-2, Athens. Vol 1, pp. 141-148.

Kilgarriff, A. and Rose, T. G. 1998. Measures for corpus similarity and homogeneity. Proc. EMNLP 3, Granada, Spain. pp. 46-52.

Roukos, S. 1996. Language Representation. In Cole (ed.) *Survey of the State of the Art in Human Language Technology*, NSF and EC. Chapter 1.6. www.cse.ogi/CSLU/HLTsurvey.html

Email: Adam.Kilgarriff@itri.brighton.ac.uk

ELRA: The European Language Resources Association

Khalid Choukri, *Chief Executive Officer, ELRA*

Since its foundation five years ago, ELRA has been working to improve language resources for the language engineering community. This report explains the methods it uses in its endeavour to meet the needs of those who use, and those who provide, language resources.

The European Language Resources Association (ELRA) was founded in 1995 as a membership association by a number of leading academic and private-sector bodies in cooperation with the European Commission. It aims to serve as a focal point for the collection, marketing, distribution, and licensing of language resources, and provides general information in the field of language engineering. ELRA is thus an efficient interface between language resource (LR) providers and LR users.

Market watch and identification of users' needs

As a provider of LRs, ELRA analyses the LR market from two points of view: that of the LR provider, by collecting LRs and proposing distribution means and solutions to legal issues; and that of the LR user, by suggesting resources that match users' needs, and related legal matters. As regards LR collection, ELRA keeps an updated inventory of existing resources, ensuring their reusability and negotiating with providers, and offers them through its catalogue. Surveys are also conducted to determine users' needs and the market segmentation of existing applications, tools, and resources. The ELRA surveys on users' needs have evolved and improved over time, and provide an excellent barometer for measuring the recent past, present, and future needs of LR users. Surveying and analysing users' needs enables us to develop a more reliable and workable business plan for our LR distribution and production activities.

Two major surveys have recently been completed: one to assess users' needs (the types of LRs that potential users work with and/or are interested in); the other to identify commercial products available. Details of the former (the 1999-2000 LR Users' Need Survey) may be seen at http://www2.echo.lu/langeng/projects/lrspp/index.html, and the preliminary results appear in the *ELRA Newsletter*, Vol. 4, No. 4. Different aspects of the survey have been presented at recent conferences and workshops, and are published in language technology newsletters and journals (Allen, 2000a-b; Allen and Choukri, 2000).

The latter survey, conducted between March and May 2000 by ELDA (the European Language Resources Distribution Agency – ELRA's distribution unit), is entitled the 'ELDA survey on multilingual issues: evolution of languages in speech and machine translation products'. It aimed to establish what commercial products are already available or planned in these domains, and to assess the evolution of the multilingual systems offered from 1995 and expected up to 2005. The results of this survey will be available at the ELRA website and in a forthcoming article in the *International Journal for Language and Documentation (IJLD)*.

Commissioning production of resources

ELRA uses the surveys to enhance its business plan and define policy for its new activity in commissioning LR production projects. It produces LR preference lists, which may give some hints about the orientations of the field. The preference lists of particular interest to the language engineering community are:

- SpeechDat-like database (1000-5000 speakers)
- Speech database for embedded systems (500-1000 speakers)
- Pronunciation lexica (for speech recognition and speech synthesis, including proper names)
- Dialogue corpus
- Enrich existing speech LRs in the ELRA catalogue
- Multilingual speech synthesis database
- Large monolingual corpora
- Parallel texts
- Bi/multilingual computational lexica
- Multimedia corpus
- Multimodal corpus

Following these preference lists and the users' need survey results, eight projects have been co-funded:

- Corpus of written Business English
- Bilingual LR dictionaries for English and Russian
- Expanding Resources for Terminology Extraction
- Italian Broadcast News Corpus
- Pronunciation lexicon of British English place names, surnames, and first names
- Scientific Corpus of Modern French
- German-French Parallel Corpus of 30 million words
- Columbian Spanish SpeechDat-like database.

One of the most recent demands for LRs is in the area of multimedia and multimodal data. Specifically, within multimodal processing, LRs are required for: face tracking; gesture recognition; facial analysis; eye-gaze tracking; face recognition; person identification; speech/lip reading; focus of attention; facial animation; and multimodal error recovery. With increasing interest in this kind of data, it is important that ELRA and ELDA continue to monitor and survey this area closely and in further detail in order to improve their response to this increasingly important area for new LRs.

Resource Quality

The reusability and quality of LR products are issues that ELRA constantly has to face. In particular, quality is ensured by a number of validation procedures. Validation criteria differ widely between speech, text and terminology data, so ELRA is establishing a network of LR validation centres.

ELRA also promotes the design and use of different LR validation manuals (currently available for written corpora and lexica, and for speech LRs), which are used for both in-production and post-production validation. It is of utmost importance that the minimum set of core criteria for LR meta-description and annotation schemas be developed. This will help to set standards to be followed by producers and expected by users, thereby ensuring the quality of LR products.

Data standardisation is also a quality criterion, with different standards relating to different types of data. Speech researchers already benefit from specified standards, as a result of an Esprit program (SAM), and the standardisation of telephone networks has contributed to that of the data they carry. For written LRs and terminology, common formats such as SGML, or the international Text Encoding Initiative (TEI), are now widespread formatting standards for the description and sharing of documents. The situation for terminology is more complex, and several projects focus on the exchange format issue (MARTIF, Geneter).

Standardisation and validation procedures according to predefined standards will contribute to both the quality and the reusability of LRs. ELRA will pursue its efforts in validation and market watch in order to offer users the very best of language resources.

References

Allen, J. 1999. Language Resources Go Digital: Update on the European Language Resources Association. In *Language International*, Vol. 11, No. 6, pp. 38-39.

Allen, J. 2000a. The ELRA/ELDA Survey of Language Resources. In *Localisation Industry Standards Association (LISA) Newsletter*. Vol. 9, No. 2, pp. 25-30.

Allen, J. 2000b The ELRA Language Resources survey: languages needed. *IJLD*, No. 5, June 2000, pp. 41-42.

Allen, J. & Choukri, K. 2000. Survey of Language Engineering Needs: a Language Resources perspective. Second International Language Resource and Evaluation Conference – LREC2000, Athens, 31 May-2 June 2000.

Choukri, K. 1998. Is there a real market for Europe's digital language resources? *Language International*, Vol. 10, No. 5, pp. 38-40.

FOR INFORMATION

ELRA initiated the Language Resources and Evaluation Conference (LREC) series. Information on this conference series can be found at: http://www.lrec-conf.org/

Khalid Choukri is Chief Executive Officer of ELRA

Email: Choukri@elda.fr
Web: http://www.elda.fr

Workshop Announcement

COMLEX 2000

Computational Lexicography and Multimedia Dictionaries

22-23 September 2000, Patras, Greece

This Workshop is organised by the Department of Electrical & Computer Engineering, University of Patras, Greece. It is supported by ISCA (International Speech Communication Association) and ELSNET

The workshop aims to present the state of the art in corpus-based monolingual and multilingual lexicography (corpora creation, methods and tools for lexical knowledge extraction, knowledge representation, etc.), and the integration of various modalities (textual, spoken, visual) in electronic dictionaries. In particular, it aims to present current work and the results of national and international projects related to the subject area.

The scientific program, to be conducted in English, will include several invited lectures as well as submitted paper presentations. The Proceedings will be available at the workshop.

Topics to be discussed at the workshop will include:

- Electronic corpora for lexicography
- Methods and tools for corpus based lexicography
- Acquisition and reusability of lexicography resources
- Annotation of lexicography resources
- Evaluation and validation of lexicography resources
- Treatment of morphology and polysemy
- Recognition of lexical units in text
- Monolingual and multilingual lexicography
- Integration of various modalities in electronic dictionaries
- Multimedia information retrieval
- Speech synthesis/recognition for multimedia dictionaries
- Encyclopaedic dictionaries
- Specialised dictionaries (language teaching, NLP applications, etc.)
- Internet based dictionaries
- Other subjects related to the workshop

The workshop will be held at the conference centre of the University of Patras, Greece. Patras is located on the west coast of Greece, 200 km from Athens, and close to the famous archaeological sites of Olympia, Delphi, and Corinth.

For further information, including registration (open until 1st September 2000), visit the COMLEX website, http://www.wcl2.ee.upatras.gr/comlex2000.htm

Talking to Computers in Bellagio

The Third Workshop on Human-Computer Conversation

Nick Webb, *University of Sheffield*

For those of you unfamiliar with Bellagio, in Northern Italy, it may be necessary to explain the five-hour lunch break. For those of us attending the Third International Workshop on Human-Computer Conversation (HCC), it was self-evident. The surroundings of the venue, the Grand Hotel Villa Serbelloni, were so beautiful that some time had to be allotted just to take it all in. The town (described in my guidebook as one of the most beautiful towns in Italy) on the shores of Lake Como is about an hour from Milan, and has played host to all previous HCC Workshops.

The spirit of these workshops is reasonably straightforward. Research in Human-Computer conversation is deemed to be something of a black art – at this time lacking both funding and real world applications. However, the fact that attendance at this year's three-day event was almost double that of previous years indicates that there is no shortage of people interested in the topic.

The Grand Hotel Villa Serbelloni, venue of the workshop

The first two invited speakers set the tone for the whole workshop. Geoffrey Leech (University of Lancaster) gave a historic overview of HCC, before discussing some theoretical findings on the grammar of English conversation. Jason Hutchens (University of Western Australia), a former Loebner prize-winner, discussed the use of purely stochastic language models to create conversational agents, a system devoid of any (linguistically motivated) theory.

These opposing views were mirrored within the second day's keynote speeches. These began with Norbert Reithinger's review of the VerbMobil project, where he extolled the virtues of large-scale language engineering projects. Following him, Harry Bunt (Tilburg University) demonstrated that Human-Human dialogue contained some purely social elements unrelated to any task, such as greetings.

David Novick (University of Texas, El Paso) brought the nature of these 'purely social elements' into question later in the day. His talk defined politeness in terms of actions of an implicit task, and stressed that politeness is used to develop and maintain a relationship, which could be of long term benefit. The matter of politeness and its role in dialogue implementations was the focus of a later discussion panel. The general feeling was that it was easy to imagine scenarios where computers needed politeness to interact with humans, but not so easy to envisage humans needing to be polite to computers. One for further investigation perhaps.

On the last day, David Traum (University of Maryland) spoke about the TRINDI project, and highlighted the gulf between theoretical and practical implementation of systems. One solution was to use dialogue toolkits, which could be used to embody some theory of dialogue processing. In contrast, the second speaker, Tomek Strzalkowski (General Electric Research Labs), spoke about practical dialogue development without reference to a specific theory, focussing on Call Centre applications of dialogue and conversation research.

The differing views of theorists versus implementers of practical systems, evident throughout the workshop, formed the focus for a lively panel discussion on the last day. It seems that those who claim to have no theory have actually embodied some theory or other, and indeed – if they were aware of this – would have greater predictive powers over the performance of their system.

Although the aim of the workshop was human-computer conversation (a purely interactional event), a number of papers were concerned with specific system dialogue (a more transactional approach). However, the focus of many of these papers was methods of making transactional systems more conversational and therefore more appealing to the user. Indeed, the prevalence of these transactional system made Marc Blasband (Compuleer) call for people to stop re-implementing train timetabling systems, unless there was some degree of novelty involved.

The workshop closed with a discussion of the next HCC workshop. Most people wanted to see even less evidence of domain specific dialogue systems, and more about general conversational systems, or the technology breakthroughs that would be required to build them. A very pleasant experience was completed with a truly excellent six-course meal at the Grand Hotel Villa Serbelloni. I advise you all to begin work on your papers for the next workshop.

FOR INFORMATION

Nick Webb is a Research Associate in the Natural Language Processing Group at the University of Sheffield.

Email: n.webb@dcs.shef.ac.uk
Web: http://www.dcsshef.ac.uk/~njw

Cross-lingual Knowledge Management: a topic coming your way

Yorick Wilks, *University of Sheffield*

Until recently, I tended to assume that knowledge management (KM) was a non-subject in the process of being 'talked up' so as to meet some imagined business need: how to organise and best access the digital information at a company's disposal with the range of tools and functions, from email to information retrieval, that research and development (R&D) workers had used daily for up to 30 years but had not had to use to run a business. This last fact is important because this R&D group (i.e., people like you and me!) are, in fact, the best 'community of practice' (as the phrase now is), concerned with how to cope with thousands of electronic files and hundreds of emails a day – problems that business is only now taking on board.

KM has been largely a matter of slogans, to do with 'the right information at the right time for the right person'. Behind this hides little more than simple text and document access techniques that build on the undoubted value of the Unix 'grep' command, as well as some degree of access to one or more classic NLP/AI functionalities, such as information retrieval/extraction (IR/IE), machine translation (MT), data mining, machine learning, summarisation, multi-modal interfaces, and the management and updating of ontologies and thesauri for browsing and searching.

Some bits and pieces of these technologies have been packaged into solutions and portals that sell widely, particularly Autonomy and Excalibur. If one examines a corporate evaluation of Automomy one reads:

Conventional technologies use keywords or frequency of associations of keywords to identify specific information. However, this approach has limitations due to the multiple definitions of some words. A much better approach is to use a concept-based methodology that understands the real meaning of words in their correct context. (Butler Group Technology Audit of Autonomy's Knowledge Suite, 1999).

This shows that the writer had no idea of the basic notions of our field. In fact, Autonomy's own Technology White Paper is no more cheering: it tells us that in *the dog came into the room, it was furry*, the computer would be 'stumped' as to whether 'it' was the dog or the room (a problem I believed to have been fully settled in 1972 or so, didn't you?), and that in *The fly, it's clear to me, can fly faster than the bee*, the computer 'may be confused by the word fly'. Or maybe not if programmed with any NLP later than the 1970s. The Automony group have a low view of NLP, and clearly no knowledge of it at all, though they have sold extensively some fairly effective but simple probabilistic pattern matching engine which operates over full text.

What is the moral here? Is it any more than, that total ignorance of a field is a good protective device for breaking into it? The important R&D question for us as a community should be: how could we now bring the best R&D to bear on KM, to give business access to the far greater power of NLP/IR/AI than is present in currently available KM packages? Perhaps the fault lies in us as an R&D community: that we have been too concerned with fragmentation and winning small competitions over a single mini-functionality, and not been syncretist enough – fusing modules for real commercial purposes. The fact that the European Commission is considering making this area a fundable one under 'Information Society Technologies' (IST) suggests there may be something in that.

But if we look at cross-lingual KM (CLKM) we see that there has been just such R&D from research groups in recent years. There is serious need there, after all, since multi-national corporations such as Unilever, Shell, and SAS, which appear to be monolingual, are often not so in practice. The usually cited need for CLKM comes from situations like that of a French engineer who urgently needs to search the files of a German colleague about a project they both worked on in, say, Spain. For more than a decade there has been not only a recognition that users need to access information in a language they do not understand, but also a brace of tools for doing this, of which the obvious public manifestation is access to MT software like Systran, attached to foreign-language WWW pages.

Cross-lingual IR and IE are now reasonably well understood, and a range of techniques have been published (for IR, in the seminal book edited by Grefenstette, and some for IE in, for example, the volume edited by Pazienza). In IE these usually take the form of some interlingual correspondence pre-established between template predicates, so that the results of searching texts in language A can be accessed in a database indexed by, or with translated data in, language B. Cross-lingual IR techniques often make use of bilingual dictionaries, or parallel multi-lingual texts, to establish the required correspondences for retrieval in the 'unknown' language, but the more adventurous are based on non-parallel texts from the same domain. As we noted above, virtually none of these techniques are yet available in any commercial form of KM.

As an R&D community, including funders and animators like the Commission, we cannot possibly predict and control where CLKM will go. The needs of journalists and publishing houses, say, may be quite different from the need to manage and access the patrimony of a large oil company. The problem of defining and meeting needs in this area is that companies themselves are still not at all sure how to manage the electronic access and communications technologies they have, and that is why simple tools like

Autonomy can thrive as they do. Many companies still seem amazed that employees use email for personal mail, or that they play games in working hours, facts lived with and accepted by the R&D community for thirty years!

References

Grefenstette, G. (ed.) 1998. *Cross-Language Information Retrieval*. Kluwer Academic.

Pazienza, M.T. (ed.) 1997. *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. Summer School, SCIE-97. Springer.

FOR INFORMATION

Yorick Wilks is Professor in the Department of Computer Science, University of Sheffield.

Email: yorick@dcs.sheffield.ac.uk

Web: http//www.dcs.shef.ac.uk/~yorick/

Report on the 3rd CLUK Colloquium

John Tait, *University of Sunderland and* ***Aline Villavicencio****, University of Cambridge*

Colloquium Report

For the first time, ELSNews reports on a new type of event – a colloquium for doctoral students. This provides an opportunity to present work in progress to, and get feedback from, peers who have not been involved in the projects. It also gives an indication of the type of work being undertaken by future researchers in the field.

This was the third in an annual series of research Colloquia organised by Computational Linguistics UK (CLUK), an informal association of people interested in computational linguistics. It provides an opportunity for doctoral students to present their work in an informal atmosphere at an early stage, and there are also some invited speakers. Although the event is restricted to work going on in the UK (with 14 papers presented and over 30 attendees from 13 different UK universities), the voices and faces came from all over the world.

The student talks covered a broad range of subject areas, and there was a marked contrast between the two days. Much of the material on the first day could be described as 'linguistics in the service of natural language engineering' (a phrase from Henry Thompson), and the following work was presented:

- Corpus-based work on discourse markers such as 'and', 'if ', 'but', etc., using relations from Rhetorical Structure Theory.
- A syntactic simplification method that emphasised the need for accurate coreference tracking to ensure text coherence.
- A natural language-based Greek Unix assistant, using recent ideas on bridging the 'generation gap' between text planning and linguistic realisation.
- A logical model to account for imperatives and actions.
- A technique to combine corpus-derived probability data with WordNet hierarchies, to improve prepositional attachment.
- An algorithm to generate deictic expressions within a document: e.g., noun phrases referring to parts of pictures.
- Text generation in different styles, based on stylistic analysis.
- Relating linguistic factors to genre analysis in a limited domain, using Biber's analysis methodology.

In contrast, the second day dealt predominantly with 'computation in the service of linguistics and cognitive science', with presentations in the following areas:

- Linear Text Segmentation, involving a method which breaks text into its smallest fragments before progressively merging 'similar' regions (the definition of 'similar' being problematic).
- Learning word order within a categorical probabilistic grammar framework.
- Learning 'semantics', interpreted (it seems) as word sense distinction and disambiguation, along with a syntax to semantics mapping – again, within a probabilistic framework. This method works surprisingly well, although the implicit assumption that there is a bounded set of possible human languages was questioned.
- An approach to learning bracketings and constituent types, using the notion that similar structures can be substituted, along with a minimum edit distance measure of similarity.
- A story comprehension system that utilises the notion of incoherence of the input text, which is contrasted with more probabilistic approaches.
- Schank's Scripts and MOPS have been resurrected in an Integrated Schema Model to story comprehension. This overcomes some problems – especially the excessive rigidity of the earlier work, although some aspects of the evaluation were considered controversial.

The event was clearly not representative of all doctoral work across the UK (with notable absences from some Universities), but it did give an indication of the wide range of good quality doctoral work going on across the UK.

FOR INFORMATION

John Tait is Professor of Intelligent Information Systems and Associate Director of Research in the School of Computing Engineering and Technology, University of Sunderland.

Email: john.tait@sunderland.ac.uk

Aline Villavicencio is a research student at the University of Cambridge Computer Laboratory.

Email: av208@cam.ac.uk

CLUK Website: http://www.dcs.shef.ac.uk/research/ilash/ CLUK/index.html

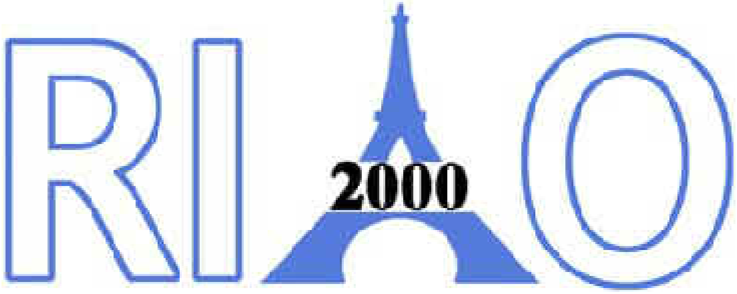
Conference Report

RIAO 2000

***Joseph Mariani** and **Donna Harman,** Co-Chairs of the RIAO 2000 Scientific Committee*

The RIAO (Computer-Assisted Information Retrieval) International Conference is held every three years. On the general theme of 'Content-Based Multimedia Information Access', the RIAO 2000 conference scope ranged from the traditional processing of text documents, to the rapidly growing fields of automatic indexing and retrieval of images and speech and, more generally, to all processing of audio-visual and multimedia information on various distribution channels, including the Net.

This year's conference, held at the Collège de France in Paris, 12-14 April 2000, was a great success. More than 400 participants from 30 different countries attended; there were 144 papers presented in 32 sessions, by authors from 26 countries; and 22 innovative applications were demonstrated.



The success was also due to the fact that scientific communities which usually meet apart were joined by the topic of the conference: Content-Based Multimedia Information Access. The information retrieval, natural language, and speech and vision communities were therefore united for the event, along with researchers from the fields of human-computer interaction and digital libraries. Furthermore, while scientific and technological issues were as usual the largest areas for presentations and discussions, the use of systems allowing humans to access multimedia information was also discussed, including human factors, sociological, and legal dimensions. Finally, as usual, RIAO 2000 exhibited its unique feature of jointly including application demonstrations and paper presentations, both selected by international programme committees.

There was a broad area of coverage: from topics related to information indexing, retrieval, routing, alerting, profiling, filtering, and summarising, to text mining and data mining, and to human-computer interaction in information retrieval and document processing. The conference emphasised the use of natural language processing in these areas, whilst extending the field to spoken language processing (including speaker and language recognition) and to image processing (including image and video indexing, browsing and retrieval, and face recognition). As well as technological issues, and issues relating to the development (architecture, best practice, standards, evaluation, resources) and use (cognitive aspects, human factors, socio-economics, security, privacy, personalisation, legal aspects) of these technologies were also considered. Special interest was given to multilingual and translingual, as well as multimedia, multimodal and transmodal processes. The applications of those technologies may be found in many areas, ranging from medical applications to business intelligence.

RIAO 2000 thereby served as a forum for cross-discipline initiatives and innovative applications. It accompanies large initiatives worldwide, such as the DARPA TIDES programme in the US, or the EC Human Language Technology and Information Filtering programmes.

Papers presented in plenary sessions addressed the extension of the domain from text to speech, vision and multimedia, the challenging areas of the relationship between the Web and the e-book, and the development of radio and TV broadcast retrieval systems. Some sessions considered multiple media, such as image and language combination, or specific ones, such as scanned document processing. New areas were considered, such as music information retrieval, while one session addressed the important issue of information visualization.

The panel sessions were the place for general discussions of hot topics, such as multilingual information access, information retrieval evaluation, and usage of information retrieval, with its legal, societal, and user dimensions. This programme was completed by two invited talks, on 'Natural Language Processing for Text Mining and Decision Making' and 'MPEG Standardisation activities – past, present and future', sponsored by ELSNET. There were also many innovative application demonstrations, covering various applications such as cross-lingual English-Arabic internet search, recognition of printed and handwritten texts, television archive retrieval, sign language indexing, machine translation, and more.

Looking at the topics of the conference (document processing; information retrieval; natural language processing; spoken language and audio processing; image and video processing; architecture, usage, and best practice and applications in different domains), the main trends are the following:

In information retrieval (IR) and document processing, we see more use of natural language processing for IR

There is a strong interest in multilingual information management, which implies the need for adequate linguistic resources, such as bilingual dictionaries and parallel texts. IR and classification techniques are combined to enhance relevance ranking and to give the user an explanation of the ranking. The interest in evaluation continues, and new metrics are proposed. There is also an increasing interest in getting access to passages of text, instead of solely documents, and a converging interest in IR and knowledge extraction – getting a direct answer to a question. The Web is of tremendous importance as a source of general information and linguistic resources; it revolutionises the whole field.

Collège de France, Paris

Natural language processing (NLP) was a major part of the conference, with 39 papers including NLP issues, and six sessions devoted to NLP. Many applications implied NLP, such as automatic indexing (keywords, phrases, parsing), question answering knowledge acquisition (thesauri, dictionaries), information extraction (templates, text mining), abstracting (from full abstracting to passage retrieval), and cross-lingual information retrieval (CLIR). There are several areas of IR to which NLP is well suited, such as automatic abstracting, question answering and CLIR, while the impact of NLP in IR remains a controversial issue, as results are mixed. NLP is a meeting point between IR and other research areas, such as video, artificial intelligence (AI) reasoning, machine translation for CLIR, shallow parsing for content access, and shallow understanding for information extraction. There are now new challenges for NLP:

- how to integrate multimodality inside NLP components (eg., layout, or image content)
- how to integrate NLP inside multimedia components (for building an interface to multimodal IR)
- how to use NLP in structural analysis
- how to use NLP in document categorisation (enriched linguistic features)
- how to combine NLP with other techniques of AI, IR, knowledge representation, data mining, etc.

In the spoken language processing domain, the fast-emerging application is audiovisual indexing. This includes automatic speech recognition on radio and TV broadcast data combined with information retrieval on the transcription, in order to conduct spoken data retrieval. Systems are usually multilingual, with experiments reported on English, French, German, Mandarin, Italian, Spanish, etc. The present recognition error rate of 20-30% is adequate for information retrieval. Browsing tools have been developed, which are now in daily use in some laboratories, and the systems are close to being ready for wide dissemination.

information from captions in video through optical character reading; musical IR; segmentation of TV data using both audio and video information. Evaluation and compilation of best practices are very active areas, as are annotation content, format, and tools, including standards such as MPEG-7, and the concept of annotation graphs.

In the field of vision processing, it appears that multimedia documents are increasingly being considered. In this conference, 77% of the papers dealt with text, compared with 11% on image, and 16% on video. Also, half of those papers dealt with image and text, while one third of the papers dealt with video, speech and text. Two approaches may be identified: signal-based and concept-based approaches; the challenge might be to combine both in order to design more advanced systems with enhanced performance. The query format and interfaces are highly specific. The applications of systems based on vision are wide, going from the very large open TV and radio broadcast indexing, browsing, and retrieval, to the specific areas of medical, biology, or land registry applications, via intermediate applications such as e-book tools. Several 'significant scale' applications were mentioned, whilst standardisation, and specifically MPEG-7, is an area of great interest.

The papers and demonstrations covered a large number of application areas. They showed very clearly that information is more than just text, and that systems must deal with various inputs: voice (from audiovisual archive or through telephone); sign language; faces; images; handwriting; video in a multilingual environment; and more – and search is becoming intelligent, just like reading.

RIAO 2000 very clearly served as a major event in the rapidly growing fields of information society technologies.

FOR INFORMATION

Joseph Mariani is, amongst other things, Director of LIMSI (and head of the Human-Machine Communication Department), Vice-President of ELRA, and a member of the ELSNET Executive Board.

Web: http://www.limsi.fr/Individu/mariani/

Donna Harman is a Group Manager at the Information Access and User Interfaces Division of the National Institute of Standards and Technology in the USA.

For further information relating to RIAO 2000
Email: riao2000@limsi.fr
Web: http://host.limsi.fr/RIAO/

Announcement of Workshop and Call for Contributions

Infrastructures for Global Collaboration

A Workshop organised by ELSNET in conjunction with ACL 2000, to be held in Hong Kong, October 7 or 8, 2000

Language and speech technologies are different from most other technologies, in that the complexity of the problems addressed is multiplied by the number of languages (every language comes with its own unique problems), and even by its square if one thinks of communication across language barriers.

Yet at the same time it is clear that solutions found for problems in one language may be fully or partially portable to other languages.

Within the European Union, the above (and other) observations have led to the creation of major R&D programmes, such as Language Engineering, and Human Language Technologies, where parties from all over Europe (and even outside) join forces in order to address the common problems. In addition, a number of infrastructures have been set up at the European level, such as ELRA (which commissions, distributes, and validates language resources), ELSNET (a network of key players in the field of Human Language Technologies, aimed at sharing information and expertise), and EAGLES/ISLE (aimed at developing standards in an international context).

Transnational infrastructures are not limited to Europe, as demonstrated by, for example, the main international professional organisations in the fields of language (ACL) and speech (ISCA) technology, and the Linguistic Data Consortium. What is lacking is a clear overview of what transnational infrastructures exist world-wide, and how they can be optimally exploited for global collaboration.

This half day workshop, with invited presentations followed by an open discussion, is aimed at people interested in R&D policies and infrastructures, and will address the following questions:

What are the existing infrastructures world-wide?

- Are they optimally exploited for global collaboration?
- If not, how could this be improved?

What infrastructures or interconnections are missing?

- What can we contribute to their creation?
- Who are the main actors (institutions, organisations)?
- What are the main instruments we have at our disposal to build and operate such infrastructures?

The intended output from the workshop is a strategic report, containing an analysis of the present situation, and an outline scenario for steps to be taken. The workshop should be seen as a first consultation and round-table discussion on this topic, to be followed by similar events at other venues, where other parts of the language and speech communities (both thematic and geographic) will be consulted.

Invitation to contribute

Everybody, both workshop participants and others, is invited to send their views on the topics addressed by this workshop to the organisers, by email. A summary of these contributions will be presented at the workshop.

FOR INFORMATION

For more information, or to contribute your views, contact Steven Krauwer, the ELSNET Co-ordinator.

Email: steven.krauwer@elsnet.org
Web: http://www.elsnet.org/acl2000workshop.html

Colloquium Digest Announcement

State of the Art in Speech Synthesis

An Institute of Electrical Engineers Colloquium

Justin Fackrell, *Lernout & Hauspie Speech Products*

This colloquium, held in London on 13 April 2000, attracted over 60 researchers working in speech synthesis – mostly from the UK, but with attendees from all over Europe and further afield. Attendees came from universities, institutes, and industry. The event was hosted by the IEE and co-sponsored by the ISCA SynSIG (Special Interest Group on Speech Synthesis), and by the Institute of Acoustics.

The 15 papers presented during the day included papers on the following areas:

- prosody – modelling and manipulation; emotion; variation
- synthesis systems – both corpus-based and parametric systems
- assessment and evaluation
- nonlinear modelling of speech signals

The programme for the event can be seen at http://www.iee.orguk/Events/e13apr00. htm.

The IEE have published the Colloquium Digest, which contains full papers of most of the presentations made at the event. For ordering information, visit http://www.iee.org.uk/publish/ordering/orderinghtml.

Future Events

Future Events in 2000

Aug 29–Sept 8 *Tbilisi Summer School in Language, Logic, and Computation*, Tbilisi, Georgia
chiko@contsys.acnet.ge http://www.geo.net.ge/llc99

Sept 5-7 *ISCA Workshop on Speech and Emotion*, Belfast, Northern Ireland
e.douglas-cowie@qub.ac.uk http://www.qub.ac.uk/en/isca/index.htm

Sept 13-16 *3rd International Workshop on Text, Speech and Dialogue (TSD 2000)*, Brno, Czech Republic
tsd2000@fi.muni.cz http://www.fi.muni.cz/tsd2000/

Sept 18-20 *ISCA ITRW International Workshop on Automatic Speech Recognition (ASR2000*), Paris, France
asr2000@limsi.fr http://www-tlp.limsi.fr/asr2000

Sept 20-23 *Architectures and Mechanisms for Language Processing (AMLaP 2000)*, Leiden, The Netherlands
AMLaP@fsw.leidenuniv.nl http://www.amlap.org

Sept 22-23 *Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, Patras, Greece
http://www.wcl2.ee.upatras.gr/comlex2000_2.htm

Sept 22-24 *5th TELRI Seminar on Corpus Linguistics*: How to Extract Meaning from Corpora, Ljubljana, Slovenia
telri-admin@ids-mannheim.de http://www.telri.de and http://nl.ijs.si/telri00/

Sept 25-28 *International Workshop on Speech and Computers (SPECOM 2000)*, St. Petersburg, Russia
specom@mail.iias.spb.ru http://www.spiiras.nw.ru/speech

Sept 26-28 *16th Conference of the Spanish Society for Natural Language Processing (SEPLN 2000)*, Vigo, Spain
sepln-secret@ei.uvigo.es http://www.coleweb.dc.fi.udc.es/sepln2000/

Oct 2-5 *Workshop on Speech Recognition and Synthesis (Prosody 2000)*, Krakow, Poland
gibbon@spectrum.uni-bielefeld.de http://www.ptfon.wmid.amu.edu.pl

Oct 3-6 *38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hong Kong
acl2k@cis.udel.edu http://www.cs.ust.hk/acl2000/

Oct 7 or 8 *The 2nd Chinese Language Processing Workshop* (in conjunction with *ACL 2000*), Hong Kong
chinese@linc.cis.upenn.edu http://www.ldc.upenn.edu/ctb/clp00.html

Oct 7 or 8 *ACL 2000 Workshop: Infrastructures for Global Collaboration*, Hong Kong
steven.krauwer@elsnet.org http://www.elsnet.org/acl2000workshop.html

Oct 16-20 *The International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China
http://www.icslp2000.org/

Oct 18-20 *Workshop on (Human-)Agent Interaction and Agent Learning*, Enschede, The Netherlands
ctwlt@cs.utwente.nl http://parlevink.cs.utwente.nl/

Oct 23-24 *International Conference on Cognitive Modelling in Linguistics 2000*, Pereslavl´-Zalesskiy (near Moscow), Russia
solovyev@mi.ru and vladimir_polyakov@yahoo.com

Nov 16-17 *22nd Conference on Translating and the Computer*, London, UK
nicole.adamides@aslib.co.uk http://www.aslib.co.uk

Nov 20-22 *Machine Translation and Multilingual Applications in the New Millennium (MT 2000)*, Exeter, UK
MT2000-request@rwsh.dircon.co.uk http://www.bcs.org.uk/siggroup/sg37.htm

Nov 23-25 *26th Annual Conference on Language Technologies*, Cologne, Germany
klaus.schmitz@fh-koeln.de http://www.fbi.fh-koeln.de/DEUTERM/ivsw2000E.htm

Dec 6-8 *The 3rd International Conference of Asian Digital Library (ADL2000)*, Seoul, Korea
info@adl2000.kaist.ac.kr

Events Coming in 2001

Jan 10-12 *The 4th International Workshop on Computational Semantics (IWCS-4)*, Tilburg, The Netherlands
Computational.Semantics@kub.nl http://cwis.kub.nl/%7Efdl/research/ti/Docs/IWCS/iwcs.htm

April 27-29 *1st International Workshop on Generative Approaches to the Lexicon (GL2001)*, Geneva, Switzerland
Pierrette.Bouillon@issco.unige.ch http://issco-www.unige.ch/conf.html

This is a selection of events – see http://www.elsnet.org/cgi-bin/elsnet/events.pl for more.

ELSNET

Office

Steven Krauwer,
Co-ordinator
Brigitte Burger,
Assistant Co-ordinator
Monique Hanrath,
Secretary
Utrecht University (NL)

Task Groups

Training & Mobility
Gerrit Bloothooft,
Utrecht University (NL)
Koenraad de Smedt,
University of Bergen (NO)

Linguistic & Speech Resources
Antonio Zampolli,
Istituto di Linguistica
Computazionale (IT)
and Ulrich Heid,
Stuttgart University (DE)

Research
Niels Ole Bernsen, NIS
Odense University (DK)
and Joseph Mariani,
LIMSI-CNRS (FR)

Industrial Panel

Harri Arnola,
Kielikone (FI)
Roberto Billi,
CSELT (IT)
Michael Carey,
Ensigma (UK)
Jean-Pierre Chanod,
Rank Xerox Research
Centre (FR)
Harald Höge,
Siemens AG (DE)
Bernard Normier,
ERLI (FR)
Brian Oakley (chair, UK)

Executive Board

Steven Krauwer,
Utrecht University (NL)
Niels Ole Bernsen, NIS,
Odense University (DK)
Björn Granström,
Royal Institute of
Technology (SE)
Nikos Fakotakis,
University of Patras (GR)
Ulrich Heid,
Stuttgart University (DE)
Joseph Mariani,
LIMSI/CNRS (FR)
José M. Pardo,
Polytechnic University of
Madrid (ES)
Geoffrey Sampson,
University of Sussex (UK)
Antonio Zampolli,
University of Pisa (IT)

ELSNET Participants

Academic Sites

AT	Austrian Research Institute for Artificial Intelligence (ÖFAI)
AT	University of Vienna
AT	Vienna University of Technology
BE	Leuven University
BE	University of Antwerp – UIA
BG	Academy of Sciences Institute of Mathematics
BY	Belorussian Academy of Sciences
CH	Istituto Dalle Molle (IDSIA)
CH	University of Geneva
CZ	Charles University
DE	Christian-Albrechts University, Kiel
DE	German Research Center for Artificial Intelligence (DFKI)
DE	Institute of Applied Information Science (IAI)
DE	Ruhr-Universität Bochum
DE	Universität Erlangen
DE	Universität Hamburg
DE	Universität Stuttgart
DE	Universität des Saarlandes
DK	Aalborg University
DK	Center for Sprogteknologi
DK	Odense University
ES	Polytechnic University of Catalonia
ES	Polytechnic University of Madrid
ES	Polytechnic University of Valencia
ES	Universitat Autonoma de Barcelona
ES	University of Granada
FR	CRIN, Nancy
FR	IRISA/ENSSAT, Lannion
FR	Inst. National Polytechnique de Grenoble
FR	Institute de Phonétique, CNRS
FR	LIMSI/CNRS, Orsay
FR	Université Paul Sabatier (Toulouse III)
GE	Tbilisi State University, Centre on Language, Logic and Speech
GR	Institute for Language & Speech Processing (ILSP), Athens
GR	NCSR 'Demokritos', Athens
GR	University of Patras
HU	Lóránd Eötvös University
HU	Technical University of Budapest
IT	Consorzio Pisa Ricerche
IT	Consiglio Nazionale delle Ricerche
IT	Fondazione Ugo Bordoni
IT	IRST, Trento
IT	Università degli Studi di Pisa
IE	Trinity College, University of Dublin
IE	University College Dublin
LT	Institute of Mathematics & Informatics
NO	Norwegian University of Science and Technology
NO	University of Bergen
NL	Eindhoven University of Technology
NL	Foundation for Speech Technology
NL	Leyden University
NL	TNO Human Factors Research Institute
NL	Tilburg University
NL	University of Amsterdam
NL	University of Groningen
NL	University of Nijmegen
NL	University of Twente
NL	Utrecht University
PT	Faculdade de Ciencias da Univ. de Lisboa
PT	INESC, Lisbon
PT	New University of Lisbon
PL	Polish Academy of Sciences
RO	Research Institute for Informatics (ICI)
SE	KTH (Royal Institute of Technology)
SE	Linköping University
RU	Russian Academy of Sciences, Moscow
UA	IRTC UNESCO/IIP, Kiev
UK	Leeds University
UK	School of Oriental and African Studies
UK	University of Science and Technology in Manchester
UK	University College London
UK	University of Brighton
UK	University of Cambridge
UK	University of Dundee
UK	University of Edinburgh
UK	University of Essex
UK	University of Sheffield
UK	University of Sunderland
UK	University of Sussex
UK	University of Ulster
UK	University of York

Industrial Sites

BE	Lernout & Hauspie Speech Products
DE	AG für Mensch-Maschine Kommunikation GmbH
DE	DaimlerChrysler AG
DE	Electronic Publishing Partners GmbH
DE	Grundig Professional Electronics GmbH
DE	IBM
DE	Langenscheidt KG
DE	Novotech GmbH
DE	Philips Research Laboratories
DE	Siemens AG
DE	Verlag Moritz Diesterweg GmbH
DE	Pc-plus computing
DK	Tele Danmark
ES	Telefonica I & D
FR	Aerospatiale
FR	LINGA s.a.r.l.
FR	LexiQuest
FR	Memodata
FR	Systran SA
FR	TGID
FR	VECSYS
FR	Xerox Research Centre Europe
FI	Kielikone Oy
FI	Nokia Research Center
GR	KNOWLEDGE S.A.
HU	MorphoLogic Ltd
IT	CSELT
IT	Olivetti RICERCA ScpA
IT	SOGEI
IT	Tecnopolis CSATA Novus Ortus
NL	Cap Gemini Nederland BV
NL	Compuleer
NL	Comsys Europe BV
NL	KPN Research
NL	Polydoc NV
SE	Telia Promotor
RU	ANALIT Ltd
RU	Russicon Company
UK	20/20 Speech Ltd
UK	ALPNET UK Limited
UK	BICC plc
UK	BT Advanced Communications Technology Centre
UK	Cambridge Algorithmica Limited
UK	Canon Research Centre Europe Ltd
UK	Ensigma
UK	Hewlett-Packard Laboratories
UK	Logica Cambridge Ltd
UK	SRI International
UK	Sharp Laboratories of Europe Ltd
UK	Vocalis Ltd

What is ELSNET?

ELSNET, the European Network of Excellence in Human Language Technologies, is funded by the European Commission's Human Language Technologies programme. Members are academic and public research institutes (85) and industrial companies (50) from all over Europe.

The long-term technological goal which unites the members of ELSNET is to build integrated multilingual natural language and speech systems with unrestricted coverage of both spoken and written language. However, the realistic prospect for commercial applications involves systems that are restricted in one way or another. Such systems are of crucial importance for Europe in that they allow implementation of, and access to, the emerging multilingual information infrastructure. These systems also contribute to the increase of European industrial competitiveness by giving better access to product and service markets across language barriers.

Building multilingual language and speech systems requires a massive joint effort by two pairs of communities: on the one hand, the natural language and speech communities, and on the other, academia and industry. Both pairs of communities are traditionally separated by wide gaps. It is ELSNET's objective to provide a platform which bridges both gaps, and to ensure that all parties are provided with optimal conditions for fruitful collaboration.

To achieve this, ELSNET has established an infrastructure for sharing knowledge, resources, problems, and solutions by offering (information) services and facilities, and by organising events which serve academia and industry in the language and speech communities.

Electronic Mailing List

elsnet-list is ELSNET's electronic mailing list. Email sent to elsnet-list@let.uu.nl is received by all member site contact persons, as well as other interested parties. This mailing list may be used to announce activities, post job openings, or discuss issues which are relevant to ELSNET. To request additions/deletions/changes of address in the mailing list, please send mail to elsnet@let.uu.nl

FOR INFORMATION

ELSNET
Utrecht Institute of Linguistics OTS, Utrecht University,
Trans 10, 3512 JK, Utrecht, The Netherlands
Tel: +31 30 253 6039
Fax: +31 30 253 6000
Email: elsnet@let.uu.nl
Web: http://www.elsnet.org